Estudio comparativo de métodos de selección de características de inferencia supervisada y no supervisada

Diego Hernán Peluffo Ordóñez

José Luis Rodríguez Sotelo

Germán Castellanos Domínguez

Resumen

En este trabajo se presenta un estudio comparativo de algunos métodos de selección de características de inferencia supervisada y no supervisada derivados del algoritmo PCA clásico. Se deduce una función objetivo de PCA a partir del error cuadrático medio de los datos y su proyección sobre una base ortonormal, y se extiende este concepto para derivar una expresión asociada al algoritmo fundamental de WPCA. Adicionalmente, se estudian los algoritmos Q - α supervisado y no supervisado y se explica su relación con PCA. Se presentan resultados empleando dos conjuntos de datos: Uno de baja dimensión para estudiar los efectos de la rotación ortogonal y la dirección de los componentes principales y otro de alta dimensión para evaluar los resultados de clasificación. Los métodos de selección de características fueron evaluados teniendo en cuenta la cantidad de características relevantes obtenidas, costo computacional y resultados

Fecha de recepción: 25 de septiembre de 2009
Fecha de aceptación: 13 de noviembre de 2009

de clasificación. La clasificación se realizó con un algoritmo particional de agrupamiento no supervisado.

Abstract

In this work, a comparative study of feature selection methods for supervised and unsupervised inference obtained from classical PCA is presented. We deduce an expression for the cost function of PCA based on the mean square error of data and its orthonormal projection, and then this concept is extended to obtain an expression for general WPCA. Additionally, we study the supervised and unsupervised $Q-\alpha$ algorithm and its relation with PCA. At the end, we present results employing two data sets: A low-dimensional data set to analyze the effects of orthonormal rotation, and a high-dimensional data set to assess the classification performance. The feature selection methods were assessed taking into account the number of relevant features, computational cost and classification performance. The classification was carried out using a partitional clustering algorithm.

Palabras Clave

Proyección ortonormal, PCA, selección de características, WPCA.

Key words

Feature selection, orthonormal projection, PCA, WPCA.

I. Introducción

En el área de reconocimiento de patrones, es común encontrarse con problemas en los que al momento de extraer patrones descriptivos de observaciones o muestras, que posteriormente serán clasificadas, no se tiene información a priori sobre la cantidad necesaria de dichos patrones, ni de la relevancia en la clasificación de los mismos. Por esta razón, los procesos de caracterización generan matrices de datos de alta dimensión, lo que puede representar un problema para la subsecuente tarea de clasificación porque puede generar bajos resultados de clasificación debido a la información redundante y además, podría implicar un costo computacional elevado. La solución a este problema se denomina selección de características o atributos.

Este problema es típico en el reconocimiento de patrones y aprendizaje de máquina y se presenta en diferentes ramas de la ciencia (procesamiento de texto, bio-informática, procesamiento de señales biomédicas, etc.).

Existen diversas alternativas para resolver esta tarea y la escogencia de un método u otro depende de las condiciones del problema y de la naturaleza de los datos (Shlens, 2009). El análisis de componentes principales (PCA) y sus variantes, como WPCA, representan una buena opción, entre otras razones, por su naturaleza no paramétrica, facilidad de implementación y versatilidad. Diversos estudios han comprobado su aplicabilidad como técnica de mapeo, extracción de características y reducción de dimensionalidad en diferentes contextos como procesamiento de señales biomédicas (Jager, 2002, Rodríguez et al, 2009), detección de rostros (Wang & Wu, 2005), entre otros.

En este trabajo se realiza un análisis comparativo de algunos métodos de selección de características de inferencia supervisada y no supervisada derivados del algoritmo PCA clásico desde el punto de vista de la proyección ortonormal y la descomposición en valores propios. Para esto se deduce una función objetivo de PCA a partir del error cuadrático medio de los datos y su proyección sobre una base ortonormal. Luego se generaliza este concepto para derivar una expresión asociada al algoritmo fundamental de WPCA. Adicionalmente, se estudia el algoritmo Q-α, propuesto en (Wolf & Shashua, 2005), para casos supervisados y no supervisados y se explica su relación con PCA. La efectividad de la selección de cada método se evalúa sobre dos conjuntos de datos, uno de baja dimensión que permite apreciar el efecto de la transformación ortonormal y la dirección de los componentes principales y otro de alta dimensión en donde la selección de características es significativa para los resultados de clasificación. Los parámetros evaluados son la cantidad de características relevantes resultantes, costo computacional y resultados de la clasificación. La etapa de clasificación se desarrolló empleando agrupamiento no supervisado de tipo particional (Rodríguez et al, 2009) con el fin de determinar si las características generan grupos homogéneos.

II. Materiales y métodos

Sea $\mathbf{x}_i$ la i-ésima observación de q atributos o características y $X = (x_i,...,x_n)^T$ la matriz de datos.

Un vector $\mathbf{x}$ de dimensión q puede escribirse como la combinación lineal de los elementos de una base ortonormal, así:

$$\mathbf{x} = \sum_{i=1}^{q} c_i \mathbf{u}_i \qquad (1)$$

donde $\mathbf{U} = (\mathbf{u}_i ,..., \mathbf{u}_q)$ representa la base ortonormal y $\mathbf{c} = (c_1 ,..., c_q)$ son los pesos de la combinación lineal. En general, cualquier proyección ortonormal $\tilde{\mathbf{x}}$ se realiza en un espacio p dimensional ($p < q$), que mejor represente a $\mathbf{x}$:

$$\mathbf{x} = \sum_{i=1}^{q} c_i \mathbf{u}_i \qquad (2)$$

El error cuadrático medio de la proyección ortonormal de la señal original y la reconstruida, puede estimarse de la forma:

$$\overline{e^2} = \varepsilon\left\{\left(\mathbf{x}-\tilde{\mathbf{x}}\right)^{\mathsf{T}}\left(\mathbf{x}-\tilde{\mathbf{x}}\right)\right\} \tag{3}$$

Reemplazando las expresiones (1) y (2) en (3), el error se puede re-escribir como:

$$\overline{e^2} = \varepsilon\left\{\left(\sum_{i=1}^{q} c_i\mathbf{u}_i - \sum_{i=1}^{p} c_i\mathbf{u}_i\right)^{\mathsf{T}}\left(\sum_{i=1}^{q} c_i\mathbf{u}_i - \sum_{i=1}^{p} c_i\mathbf{u}_i\right)\right\} = \left\{\left(\sum_{i=p+1}^{q} c_i\mathbf{u}_i\right)^{\mathsf{T}}\left(\sum_{i=p+1}^{q} c_i\mathbf{u}_i\right)\right\} \tag{4}$$

A. Análisis de componentes principales

En estadística, la aplicación más común de PCA es la reducción de la dimensionalidad de un conjunto de datos. La idea general de este método es determinar el número de elementos descriptivos subyacentes tras un conjunto de datos que contengan información de la variabilidad de dichos datos (Shlens, 2009). En otras palabras, en PCA se busca la proyección en la que los datos queden mejor representados en términos de mínimos cuadrados, dicha proyección corresponde a la varianza acumulada de cada observación. Una de las ventajas de PCA es que reduce la dimensionalidad de un conjunto de datos, reteniendo aquellos atributos o características del conjunto de datos que contribuyen más a su varianza, por tanto las características escogidas son las que presentan mayor separabilidad con respecto a la media de los datos.

PCA construye una transformación lineal de los datos originales de manera que se genere un nuevo sistema de coordenadas en donde la mayor varianza del conjunto de datos es capturada en el primer eje (denominado primer componente principal), la segunda varianza más grande en el segundo eje, y así sucesivamente; donde la medida de varianza la define una estimación de la matriz de covarianza de los datos (Shlens, 2009). Por tanto, el objetivo de PCA es minimizar el error cuadrático medio de la proyección de los datos sobre los vectores propios de la matriz de covarianza, sujeto a una condición de ortonormalidad. Minimizar dicho error (ecuación (4)), es equivalente a maximizar el complemento del mismo, es decir:

$$\varepsilon\left\{\left(\sum_{i=1}^{p} c_i \mathbf{u}_i\right)^{\mathsf{T}}\left(\sum_{i=1}^{p} c_i \mathbf{u}_i\right)\right\} = \varepsilon\left\{\left(\sum_{i=1}^{p} c_i^{2}\right)\right\} = \mathrm{tr}\left(\mathbf{c}_p{}^{\mathsf{T}}\mathbf{c}_p\right) = \mathrm{tr}\left(\mathbf{C}_p\right) \quad (5)$$

donde $\mathbf{c}_p$ es un vector compuesto por los p primeros elementos de $\mathbf{c}$ y tr(∑) denota la traza de su matriz argumento.

Se puede apreciar que $\mathbf{c}_p$ es una matriz simétrica y semipositiva definida. En el caso de PCA, para realizar la proyección de todos los vectores $\mathbf{x}_i$, la matriz $\mathbf{C}$ es de q x q y corresponde a la matriz de covarianza que puede ser estimada como:

$$\mathbf{C}_{PCA} = \frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{X} \quad (6)$$

La anterior ecuación se aplica después de centrar los datos en la media de cada observación, es decir:

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \mu\left(\mathbf{x}_i\right),\ i=1,\ ...,n \quad (7)$$

donde μ(∑) representa la media.

Con lo anterior y considerando el criterio de invariancia ortonormal (Yu & Shi, 2003), puede plantearse el siguiente problema de optimización:

$$\max \frac{1}{n}\mathrm{tr}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{U}\right) = \sum_{j=1}^{p}\lambda_j \quad (8)$$

donde $\mathbf{I}_d$ es una matriz identidad de dimensión d y λ_j on los valores propios de $\mathbf{C}_{PCA}$.

$$\text{s. t. } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_p \quad (9)$$

Debido a la simetría del matriz $\mathbf{C}_{PCA}$, existe una base completa de vectores propios de la misma y por tanto la transformación lineal que mapea los datos a esta base es, justamente, la representación de los datos que se utiliza para la reducción de la dimensionalidad. Los elementos de la base ortogonal se denominan componentes principales y la proyección de los datos se obtiene con:

$$\mathbf{Z}_{PCA} = \mathbf{XU} \tag{10}$$

En la ecuación (8) se aprecia que el valor de la función objetivo se asocia directamente a la suma de los valores propios de la matriz de covarianza, por tanto la solución de de este problema de optimización conduce al absurdo de tomar todas las características, por esta razón es necesario aplicar un criterio de adicional sobre los componentes principales para llevar a cabo la selección de características.

En el algoritmo general de PCA, se considera $\mathbf{U}$ como la matriz de vectores propios de $\mathbf{C}_{PCA}$ordenados de forma descendente, esto es

$$\begin{gathered}[\mathbf{U}\mathbf{\Lambda},] = \mathrm{eig}(\mathbf{C}_{PCA}),\ \mathbf{\Lambda} = \mathrm{Diag}(\boldsymbol{\lambda}), \\ \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q) \text{ y } \mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_q) \text{ con } \lambda_1 > \ldots > \lambda_q\end{gathered} \tag{11}$$

donde eig (λ) representa la descomposición en valores y vectores propios, y Diag (λ) denota la matriz una matriz diagonal formada por el vector de su argumento.

Por último se escogen los p primeros componentes principales como los elementos relevantes, es decir, los que mejor representan a $\mathbf{X}$ en términos del error cuadrático medio. El valor de p, puede definirse a través de algún criterio de varianza acumulada, o evaluando iterativamente los resultados de un clasificador, escogiéndose al final los elementos que presenten un error de clasificación admisible. El criterio de varianza acumulada se aplica sobre el valor de la función objetivo normalizado, así:

$$\mathbf{z}^* = \frac{\mathrm{diag}\left(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X}\mathbf{U}\right)}{\mathrm{tr}\left(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X}\mathbf{U}\right)} = \frac{\mathrm{diag}(\mathbf{\Lambda})}{\mathrm{tr}(\mathbf{\Lambda})} \tag{12}$$

$\mathbf{z}^*$ es un indicador de la variabilidad de cada componente, entonces, para un criterio del $N\%$ se consideran los p elementos de $\mathbf{Z}_{PCA}$ que correspondan a un valor de varianza acumulada del $N\%$, es decir, $\sum_{i=1}^{p} z_i^* \approx N/100$.

B. PCA ponderado

La diferencia de este método y el algoritmo básico de PCA radica en la estimación de la matriz de covarianza. En PCA ponderado (WPCA, de sus siglas en inglés), se emplea una matriz de covarianza denominada ponderada, para la cual existen dos maneras fundamentales de estimación (Yue & Tomoyasu, 2003). La primera corresponde al caso de la ponderación por características, es decir:

$$\mathbf{C}_{WPCA} = \mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{W} \qquad (13)$$

donde **W**=Diag (**w**) es una matriz diagonal de pesos. De este modo se puede cambiar la importancia relativa de las características en la representación de **X**.

Por tanto, el problema de optimización para WPCA con ponderación de características podría plantearse como:

$$\max \operatorname{tr}\left(\mathbf{U}^{\top}\mathbf{W}^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{W}\mathbf{U}\right) = \sum_{j=1}^{p} \lambda_j \qquad (14)$$

donde λ_j son los valores propios de $\mathbf{X}^{\top}\mathbf{X}$.

En este caso, la proyección sobre los componentes principales se obtiene con: $\mathbf{Z}_{WPCA} = \mathbf{XWU}$.

La segunda forma básica de WPCA consiste en la ponderación de las observaciones o muestras, donde la matriz de covarianza se puede escribir como:

$$\mathbf{C}^{s}_{WPCA} = \mathbf{X}^{\top}\mathbf{W}^{\top}\mathbf{W}\mathbf{X} \qquad (15)$$

Nótese, que estimar la covarianza ponderada empleando (16), es equivalente a ponderar las observaciones y luego calcular la matriz de covarianza usando (6). En efecto, si $\mathbf{X}_w$=$\mathbf{WX}$ representa los datos ponderados, entonces la proyección se realiza sobres los valores propios de $\mathbf{X}_w{}^{\mathsf{T}}\mathbf{X}_w$ que corresponde a $\mathbf{C}^{s}_{WPCA}$.

El problema de optimización de WPCA empleando ponderación de observaciones es:

$$\max \operatorname{tr}\left(\mathbf{U}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{W}\mathbf{X}\mathbf{U}\right) = \sum_{j=1}^{p} \lambda_j \qquad (17)$$

$$\text{s. t. } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_p,\ \mathbf{W}^{\mathsf{T}}\mathbf{W} = \mathbf{I}_p \qquad (18)$$

Con esto, la proyección sobre los componentes principales se determina con $\mathbf{Z}^{s}_{WPCA} = \mathbf{XUW}$. Existen diversas formas de estimar los pesos. En la Tabla 1 se muestran algunos métodos no supervisados de ponderación.

Método	Expresión
Pre-normalización de PCA	$w_i = 1/\sqrt{\frac{1}{q}\sum_{j=1}^{q} X_{ij}^2}$
Ponderación empleando los valores propios (Wang & Wu, 2005)	$w_i = \lambda_i^{-1/2}$

Tabla 1. Algunos métodos de ponderación no supervisada de WPCA

También existen métodos supervisados en los que se consideran las etiquetas para la definición de los pesos.

En general, la ponderación w_i se aplica partiendo de la idea de que no siempre los elementos que generan una buena representación, generan también una buena separabilidad (Wang & Wu, 2005). Además, un elemento relevante en la representación, deberá seguir siendo relevante pese a la ponderación; mientras que un elemento que no sea relevante en la representación pero que si lo sea en la clasificación, podría llegar a ser considerado dentro del conjunto de los componentes principales después de dicha ponderación.

C. Algoritmo Q-α

En (Wolf & Shashua, 2005) se presenta una definición de relevancia en términos de la matriz de "afinidad" que captura los productos internos de las observaciones y un vector de ponderación. Este concepto se basa en la coherencia de los subconjuntos o *clústeres* resultantes de un proceso de agrupamiento, empleando propiedades espectrales y análisis topológico derivado de la teoría de grafos, donde la matriz de datos representa los vértices de los grafos no dirigidos y la matriz de afinidad indica los pesos de cada arista del grafo (Yu & Jiambo, 2003). A este método, los autores lo denominaron *Q-α* debido a que *α* es el vector de ponderación y ***Q*** es la matriz ortonormal de rotación. En este estudio, la matriz ortonormal se denota con **U**.

Sea **M** una matriz de q x n, definida como $\mathbf{M} = \mathbf{X}^{\top} = (\mathbf{m}_1, \ldots, \mathbf{m}_q)^{\top}$ y pre-procesada de forma que los vectores $\mathbf{m}_i$ tengan media cero y norma unitaria, entonces la matriz de afinidad se puede obtener con:

$$\mathbf{C}_{\boldsymbol{\alpha}} = \sum_{i=1}^{q} \alpha_i \mathbf{m}_i \mathbf{m}_i^{\top} = \mathbf{M}^{\top} \operatorname{Diag}(\boldsymbol{\alpha})\, \mathbf{M} \qquad (19)$$

Desde el punto de vista estadístico, la idea general del algoritmo *Q-α* es la misma de PCA, parte de un principio de variabilidad acumulada basada en mínimos cuadrados, con la diferencia que, en este caso, la variabilidad se mide con la matriz $\mathbf{C}_{\mathbf{á}}$ y no con la covarianza. El factor de escalamiento α_i permite ajustar la importancia relativa de cada característica. Intuitivamente, a partir de estas premisas se podría plantear una función objetivo como la mostrada en (8), sin embargo, dado que en este algoritmo se parte de una transformación ortonormal arbitraria, **U** no representa los componentes principales y por tanto la selección de características se relaciona directamente con el valor de α_i. Entonces, considerando que la solución del problema consiste en encontrar el vector **á** , resulta conveniente re-formular el problema de optimización planteado en (8) a la siguiente forma cuadrática:

$$\max_{\boldsymbol{\alpha},\,\mathbf{U}} \operatorname{tr}\left(\mathbf{U}^{\mathsf{T}}\mathbf{C}_{\boldsymbol{\alpha}}\mathbf{C}_{\boldsymbol{\alpha}}\mathbf{U}\right)=\sum_{j=1}^{p}\lambda_j \qquad (20)$$

$$\text{s. t. } \mathbf{U}^{\mathsf{T}}\mathbf{U}=\mathbf{I}_p, \boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{\alpha}=1 \qquad (20)$$

donde λ_j son los valores propios de $\mathbf{C}_{\mathbf{á}}$ y $\mathbf{U}$ es una matriz ortonormal.

Duplicar $\mathbf{C}_{\mathbf{á}}$ no cambia la naturaleza de la función objetivo porque este término es una matriz simétrica y semipositiva definida, además, en términos matemáticos, esto resulta ventajoso porque permite plantear una forma cuadrática con respecto a la variable de interés, así:

$$\max_{\mathbf{á}} \mathbf{á}^{\mathsf{T}}\mathbf{G}\mathbf{á} \qquad (20)$$

$$\text{s. t. } \mathbf{á}^{\mathsf{T}}\mathbf{á}=1 \qquad (20)$$

donde $\mathbf{G}$ es una matriz auxiliar cuyas componentes son $G_{ij}=\left(\mathbf{m}_i^{\mathsf{T}}\mathbf{m}_j\right)\mathbf{m}_i^{\mathsf{T}}\mathbf{U}\mathbf{U}^{\mathsf{T}}\mathbf{m}_j$.

Estas últimas ecuaciones corresponden al problema de optimización de la versión no supervisada de *Q-α*. Dado que, en principio, la matriz **G** es obtenida a partir de una transformación ortonormal arbitraria, es necesario plantear un método iterativo en el que se sintonicen la matriz **U** y el vector **á**. Del problema de optimización planteado en (20), se aprecia que mientras el vector **á** apunta a la dirección de las características relevantes, **U** indica su rotación; por tanto la sintonización de estos parámetros es mutuamente dependiente y debe realizarse de forma alternante como se muestra en el siguiente algoritmo:

1. Inicialización: $\mathbf{M}=\mathbf{X}^{\mathsf{T}}$, $\mathbf{U}^{(0)}$ de k x n ($\mathbf{U}^{(0)\mathsf{T}}\,\mathbf{U}^{(0)}=\mathbf{I}_n$), $\mathbf{m}_i \leftarrow (\mathbf{m}_i - \mu(\mathbf{m}_i)) \,/\, \|\mathbf{m}_i\|$
2. Formar $\mathbf{G}$: $G_{ij}=(\mathbf{m}_i^{\mathsf{T}}\mathbf{m}_j)\,\mathbf{m}_j^{\mathsf{T}}\,\mathbf{U}\mathbf{U}^{\mathsf{T}}\mathbf{m}_j$

3. Calcular **á** como el vector propio asociado al máximo valor propio de **G**
4. Calcular $\mathbf{C}_á$: $\mathbf{C}_á\, \mathbf{M}^T \mathrm{Diag}\,(\boldsymbol{\alpha})\mathbf{M}$
5. Obtener la transformación ortonormal de $\mathbf{C}_á$: $\mathbf{Z}^{(r)} = \mathbf{C}_á^{(r)}\, \mathbf{U}^{(r\text{-}1)}$
6. Descomposición *QR* de $\mathbf{Z}^{(r)}$: $[\mathbf{U}^{(r)}, \mathbf{R}] = \mathrm{qr}(\mathbf{Z}^{(r)})$
7. Incrementar r: $r \leftarrow r + 1$ y retornar al paso 2

En (Wolf & Shashua, 2005), también se estudian dos alternativas del Q-α no supervisado: La primera, empleando una normalización a través del laplaciano de la matriz de afinidad, y la segunda, basada en el criterio de la aceleración de Ritz y descomposición en valores singulares cuando se asume un vector **á** inicial. Adicionalmente, se demuestra que la convergencia del algoritmo ocurre en las primeras 4 iteraciones. No obstante, un indicador de la convergencia del algoritmo puede establecerse con el cambio del vector **á**, es decir, la diferencia del vector obtenido en la iteración actual y el obtenido en la iteración inmediatamente anterior: $\|\mathbf{á}^{(r)} - \mathbf{á}^{(r\text{-}1)}\| < \delta$, donde δ define la precisión.

En el paso 1 del algoritmo de Q-α no supervisado, se establece una matriz ortonormal inicial de n x k, donde k: $k < n$ puede ser arbitrario, dado que no tiene implicación en la función objetivo a razón de que esta matriz se convierte en una matriz cuadrada de n x n a partir de la segunda iteración.

Es fácil comprobar que la solución del problema de optimización planteado en (22) corresponde al vector propio asociado al mayor valor propio de **G**. De la definición de valores y vectores propios se tiene que:

$$\mathbf{G}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \Rightarrow \boldsymbol{\alpha}^{-1}\,\mathbf{G}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{T}\mathbf{G}\boldsymbol{\alpha} = \lambda \qquad (24)$$

Por tanto, para que $\mathbf{á}^T\mathbf{G}\mathbf{á}$ sea máximo, **á** debe estar asociado al mayor valor propio de **G**.

En los pasos 5 y 6, se realiza una proyección ortonormal de $\mathbf{C}_\mathbf{a}$ y se aplica descomposición *QR*, respectivamente, para obtener la matriz **U** refinada para la siguiente iteración. Por último, las p características relevantes se seleccionan como los elementos de $\mathbf{M}$ (ponderada por el vector $\boldsymbol{\alpha}$ resultante) que cumplan $\sum_{i=1}^{p} \alpha_i^2 \approx N/100$ para un criterio del N %.

Q-α supervisado: En la versión supervisada del algoritmo *Q-α*, la matriz $\mathbf{C}$ puede indicar la afinidad intra y entre clases empleando etiquetas, es decir:

$$\mathbf{C}_{\boldsymbol{\alpha}}^{gh} = \sum_{i=1}^{q} \alpha_i \mathbf{m}_i^g \mathbf{m}_i^{h\top} \tag{25}$$

donde g y h son indicadores de las clases.

De acuerdo a esta definición puede plantearse el siguiente problema de optimización:

$$\max_{\boldsymbol{\alpha}, \mathbf{U}^{gh}} \sum_{l=1}^{k} \operatorname{tr}\left(\mathbf{U}^{ll\,\top} \mathbf{C}_{\boldsymbol{\alpha}}^{ll} \mathbf{C}_{\boldsymbol{\alpha}}^{ll} \mathbf{U}^{ll}\right) - \gamma \sum_{g \neq h} \operatorname{tr}\left(\mathbf{U}^{gh\,\top} \mathbf{C}_{\boldsymbol{\alpha}}^{gh} \mathbf{C}_{\boldsymbol{\alpha}}^{gh} \mathbf{U}^{gh}\right) \tag{26}$$

$$\text{s. t. } \mathbf{U}^{gh\,\top} \mathbf{U}^{gh} = \mathbf{I}_p, \ \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1 \tag{27}$$

La anterior formulación captura los dos objetivos que generan un buen agrupamiento, desde el punto de vista de topológico: Maximiza la afinidad intra-clase (parte izquierda de la función) y minimiza la afinidad entre clases (parte derecha de la función), todo sujeto a una condición de ortonormalidad. El parámetro de regularización γ define la prioridad de lo que se desea maximizar en la función objetivo (parte derecha o izquierda) y su valor debe encontrarse en el rango $0 < \gamma < 1$.

Este problema también puede interpretarse como:

$$\max_{\boldsymbol{\alpha}} \sum_{l=1}^{k} \boldsymbol{\alpha}^\top \mathbf{G}^{ll} \boldsymbol{\alpha} - \gamma \sum_{g \neq h} \boldsymbol{\alpha}^\top \mathbf{G}^{gh} \boldsymbol{\alpha} = \max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \mathbb{G} \boldsymbol{\alpha} \tag{28}$$

$$\text{s. t. } \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = 1$$

donde $G_{ij}^{gh} = \left(\mathbf{m}_i^{g\top} \mathbf{m}_j^g\right) \mathbf{m}_i^{h\top} \mathbf{U}^{gh} \mathbf{U}^{gh\top} \mathbf{m}_j^h$ y $\mathbb{G} = \sum_{l=1}^{k} \mathbf{G}^{ll} - \gamma \sum_{g \neq h} \mathbf{G}^{gh}$.

De este modo, se puede aplicar directamente el algoritmo Q-α no supervisado, usando **G** como matriz auxiliar.

III. Resultados y discusión

para la evaluación de los métodos de selección de características se emplearon dos conjuntos de datos. El primer conjunto de datos (DS1) es de 3 clases, 150 observaciones y 3 características y fue utilizado para estudiar los efectos de la transformación ortonormal y la dirección de los componentes principales. Este conjunto de datos fue generado artificialmente con funciones de densidad de probabilidad normal multivariada $N(\mu,\sum)$, donde las clases se definen con el valor de la media y una matriz de covarianza aleatoria. La distancia entre medias, que corresponde a la distancia entre clases, se fijó en 10. El conjunto de datos fue diseñado de tal forma que existan 50 elementos por cada clase.

El segundo conjunto de datos (DS2) es de 3 clases, 2200 observaciones y 26 características, y corresponde al registro 207 de la base de datos MIT/BIH (Moody & Mark, 1999), el cual contiene 4 clases desbalanceadas y de alta variabilidad. El registro fue caracterizado como se explica en (Rodríguez et al, 2009). El conjunto de datos DS2 se utilizó para evaluar la cantidad de características relevantes y los resultados de la clasificación considerando clases desequilibradas.

La clasificación se realizó con un algoritmo de *clustering* particional basado en densidades, descrito en (Rodríguez et al, 2009), y fue evaluada en términos de especificidad (Sp) y sensibilidad (Se) por cada clase, además, se tuvieron en cuenta otros factores como el tiempo y la cantidad de características relevantes obtenidas (p). También se aplicó una medida semi-supervisada f_1/f_2 que representa la relación del valor óptimo de la función objetivo (f_1) y el calculado con el agrupamiento resultante (f_2), esta medida indica que un agrupamiento se hizo correctamente cuando su valor es cercano a 1. En (Rodríguez et al, 2009) se encuentra la descripción detallada de esta medida. Para llevar a cabo la etapa de clasificación, se estableció como número de clústeres 8, porque esta cantidad es razonable en comparación con el número de total de clases presentes en DS2 y además, ha registrado buen desempeño (Rodríguez et al, 2009).

Las pruebas se realizaron con MatLab R2008a, en un computador de 2 de GB de RAM y un procesador Core 2 Quad de 2.8 Ghz.

En la Figura 1 se muestra la representación resultante de la transformación lineal de cada uno de los métodos estudiados, donde cada color representa una clase.

La Figura 2 muestra los valores de relevancia de cada una de las características de DS2 aplicando los diferentes métodos de selección de características. En el caso de PCA y WPCA, la relevancia corresponde al índice de varianza acumulada mostrado en la ecuación (12) y en *Q-α* corresponde al vector **á**. En el algoritmo *Q-α* supervisado, el parámetro de regularización se fijó en $\gamma = 0.5$. Los resultados de clasificación se muestran en la Tabla II.

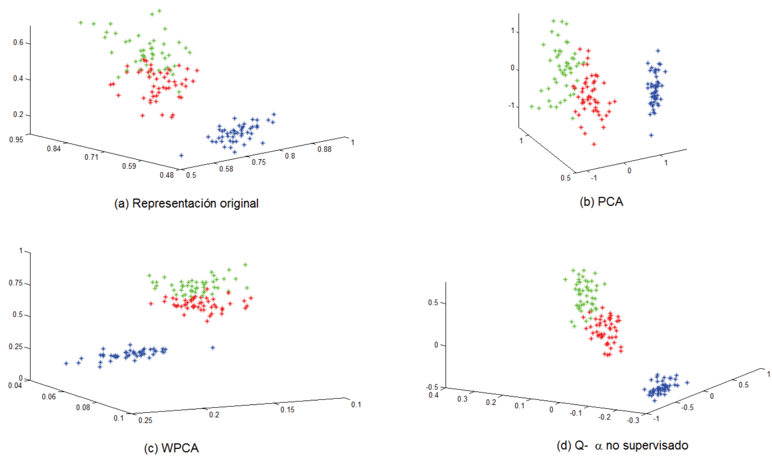
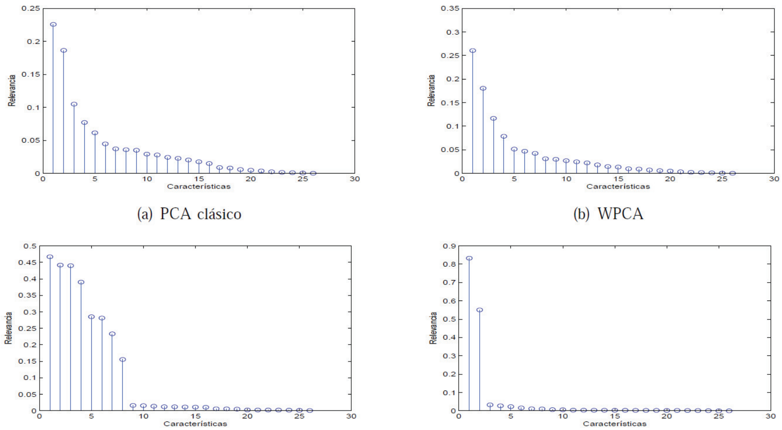


Figura 1. Representación de las características de DS1 empleando los diferentes métodos



(a) PCA clásico (b) WPCA (c) $Q-\alpha$ supervisado (d) $Q-\alpha$ no supervisado

Figura 2. Valores de relevancia de DS2 aplicando los métodos estudiados

Tabla 2. Resultados obtenidos con DS2 aplicando selección de características y *clustering* particional

Método	**Clases**	**Clase 1**	**Clase 2**	**Clase 3**	**Tiempo**	f_1/f_2	***p***
	Elementos por clase	**1542**	**106**	**210**	**(s)**		
PCA	Se Sp	93.77 94.03	98.1 93.4	95.24 99.76	49.54	0.94	10
WPCA	Se Sp	93.77 97.17	98.1 94.58	95.24 99.76	47.96	0.96	9
Q-α **no supervisado**	Se Sp	97.41 98.1	98.1 98.29	96.67 99.21	46.67	0.98	2
Q-α **supervisado**	Se Sp	98.83 99.22	99.05 98.69	98.91 99.94	46.09	0.99	8

En la Figura 1 se aprecia que la transformación lineal resultante de cada método genera una representación de los datos en donde quedan mejor representados en términos de la variabilidad medida a través de mínimos cuadrados. En el caso de WPCA, aplicando ponderación con valores propios (ver Tabla 1), se aprecia la relación directa del efecto de la ponderación con la separabilidad de las clases. El método *Q-α* no supervisado es el más estricto o "fuerte" al momento de seleccionar las características, como se aprecia en la Figura 2 y la Tabla 2, esto se debe a que la relevancia la mide el primer (que es también el mayor) vector propio de la matriz auxiliar y por esta razón la variabilidad tiende a concentrarse radicalmente en los primeros elementos.

En general, todos los métodos exhibieron buenos resultados. En este caso los que registraron mejor desempeño fueron los algoritmos *Q-α* (ver Tabla 2), porque son menos sensibles a las clases desbalanceadas. Además, registraron menor costo computacional.

También se puede apreciar que *Q-α* supervisado implicó menos tiempo de procesamiento que la versión no supervisada, esto se debe a que el análisis de las submatrices auxiliares tardó menos que el análisis de toda la matriz auxiliar. Sin embargo, vale advertir que esto no es una constante, en algunos casos podría darse que tarde menos tomar toda la matriz que realizar el análisis por submatrices.

IV. Conclusiones

La selección de características es, en muchas ocasiones, una etapa imprescindible en el diseño de los sistemas de clasificación y reconocimiento de patrones, sin embargo, la tarea de escoger un método no es trivial, existen diversos factores a tener en cuenta como costo computacional, conocimiento del conjunto de datos y objetivo de la clasificación y todos ellos enmarcados en los detalles particulares del problema que se vaya a resolver. En este trabajo se presentaron algunas alternativas para la selección de características (PCA clásico, WPCA y *Q-α*), con su correspondiente deducción matemática y aplicación.

Con en este estudio se logró plantear una nueva perspectiva del análisis de componentes principales aplicando principios de ortonormalidad, de donde se infirió un problema de optimización genérico que pudo ser extendido a variantes como WPCA y permitió explicar el algoritmo *Q-α* en su versión supervisada y no supervisada. Además, se justificó el concepto de variabilidad acumulada en la selección de características y se desarrollaron criterios de selección para cada uno de los métodos estudiados.

V. Agradecimientos

Los autores agradecen al programa de beca para estudiantes sobresalientes de posgrado de la Universidad Nacional de Colombia y al programa de financiación para Doctorados Nacionales de Colciencias. Así como también, al grupo de trabajo académico "Grupo de Control y Procesamiento Digital de Señales (GC&PDS)" de la Universidad Nacional de Colombia – sede Manizales.

Bibliografía

Jager, F. (2002). Feature extraction and shape representation of ambulatory electrocardiogram using the karhunen-lòeve transform. *Electrotechnical Review* 69 (2), 83–89.

Moody, G.B. & Mark, R.G. (1999). The MIT-BIH Arrhytmia Database on CD-ROM and software for use with it. *Computers in Cardiology, CINC.*

Rodríguez, J.L., Peluffo, D., Cuesta, D. & Castellanos, G. (2009). Non-parametric density-based clustering for cardiac arrhythmia analysis. *Computers in cardiology, CINC.*

Shlens, J. (2009). A tutorial on principal component analysis.

Wang, H.Y. & Wu, X.J. (2005). Weighted pca space and its application in face recognition. *Machine Learning and Cybernetics*, 2005. *Proceedings of 2005 International Conference on. Vol. 7.*

Wolf, L. & Shashua, A. (2005). Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. *Journal of machine learning.* 6, 1855 – 1887.

Yu, S.X. & Shi, J. (2003). Multiclass spectral clustering. *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision.* IEEE Computer Society, Washington, DC, USA, p. 313.

Yue, Y.H. & Tomoyasu, M. (2004). Weighted principal component analysis and its applications to improve fdc performance. *Conference on decision and control.*